Focused crawling in the ALVIS semantic search engine

Anders Ardö
KnowLib research group, Department of Information Technology
Lund University, P.O. Box 118, SE-221 00 Lund, Sweden
Anders.Ardo@it.lth.se

ABSTRACT

The EU project ALVIS - Superpeer Semantic Search Engine, aiming at developing an Open Source prototype of a peer-to-peer, semantic based search engine, is briefly presented. A focused (or topic specific) crawler, responsible for creating topic-specific databases within ALVIS, is presented in more detail. It is based on a combination of a standard Web crawler and an automated subject classifier. The topic focus is provided by an ontology that is used as topic definition. When a document have been deemed relevant further processing (like character set normalization, language identification and simple text segmentation), is done in preparation for the ALVIS processing pipeline.

1. INTRODUCTION

The vast quantity of information sets new challenges for even the best commercial search engines. Intelligent searching of the Internet is not just a question of scaling existing techniques. What is needed is a departure from the keyword search that has made current search technology cumbersome even for the skilled. Qualitatively better ways are needed to allow more meaningful, semantically processed queries. This is a problem involving natural language processing, information extraction and general artificial intelligence.

ALVIS [1] conducts research in the design, use and interoperability of topic-specific search engines with the goal of developing an Open Source prototype [6] of a distributed, semantic-based search engine. Distribution is intended to be able to work with heterogeneous search servers, using query topics as a routing mechanism, and using distributed methods for ranking.

2. ALVIS

ALVIS approach is not the traditional Semantic Web approach with coded meta-data, but rather an engine that can build on content through semi-automatic analysis. However focused crawling depend on pre-existing topic definition ontologies and uses coded meta-data from crawled Web-pages in order to keep only relevant documents. Natural language processing, statistical processing and a probabilistic document model of the corpus extracts semantic information from both queries and documents thus providing disambiguation, semantic clarification and topical content that can enrich both indexed documents and queries.

To demonstrate the scalability of a distributed peer-to-peer search system new basic algorithms and protocols for distributed search needs to be developed. This includes efficient query distribution and result merging using implicit and automatically generated semantics.

The combination of design, distributed operation and Open Source development have been chosen to support incremental growth, third-party involvement, low barrier to entry, as well as provide a small degradation of quality in results over an equivalent monolithic system.

2.1 ALVIS architecture

The ALVIS document processing system takes input from a source, for example a focused crawler. The internal document representation is a XML-structure that contains the original document information as well as enriched data added in the processing pipeline. Using a canonical XML format makes uniform document handling easier in the processing pipeline.

Most stages in the document processing pipeline enriches the document representation with additional information, for example by linguistic processing (lemmatization, part-of-speech tagging, named entity extraction, etc) and probabilistic modeling (ranking, categorization).

Distributed information retrieval is handled by a layered P2P architecture [3] that provides functionality for distributed indexing and query/retrieval as well as shared ranking.

3. FOCUSED CRAWLING

Focused Web-crawling is an integral component of ALVIS, capable of generating topic-specific databases of Web-pages by crawling the Web and only save relevant (i.e. topic-specific) pages. The crawler uses techniques for automated subject classification developed within the KnowLib research group [2] in order to stay focused. The software is freely available from [4].

The key to a successful focused Web crawler is to enable the crawler to select the most relevant links to follow in order to find the most relevant (with respect to the focus topic) pages. This is done by evaluating the relevance of a page for a specific topic (or subject) area (for example by automated subject classification). Various independent

Demos and Posters of the *2nd European Semantic Web Conference (ESWC 2005)*, Heraklion, Greece, 29. May - 1. June 2005
.

methods will be studied in order to improve and validate the classification assigned by the automated classification component of the focused crawler.

Ranking of crawled resources is essential for efficient focused crawling. It can utilize various methods like the traditional tf/idf ranking, PageRank, or scores based on occurrence of semantically important terms. Ranking algorithms can be both local (using only information from one page) and global (using information from the entire database) in scope. To improve the overall ranking global methods based on clustering (utilizing linking structures, co-linking, statistical methods using term-frequencies) as well as score inheritance through links (topical PageRank) can to be taken into account. The possibility to use distributed ranking will be studied to improve rankings when the focused crawler is part of a distributed search system. This also ensures that a focused crawler can contribute to distributed ranking calculation. Algorithms for efficient scheduling of URLs for crawling based on ranking of the page(s) the linking to the specific URL will be developed.

Almost all above methods rely on the database being reasonably clean and free from unnecessary noise. Among other aspects there is a need to incorporate efficient methods and tools for detecting and handling server name aliases as well as detection of almost identical pages. Identical pages are trivial to detect. But almost identical is a harder problem. Differences can be trivial changes like different backgrounds or images, visit counters or other insignificant text changes.

3.1 Ontology based automated classification

The system for automated subject classification, that determines topical relevance, is based on matching of terms from a topic definition with the text of the document to be classified ([5]). The topic definition forms an ontology with subject classes in a hierarchical classification system and terms associated with each subject class. Terms can be single words, phrases (ie a number of words in exact order), or boolean AND-expressions connecting terms (ie all terms must be present but in any order). Boolean OR-expressions are implicitly handled by having several different terms associated with the same subject class.

Each time a match is found the document is awarded points based on which term is matched and in which structural part of the document (loc) the match is found. The points are summed to make the final score of the document. If that score is above a cut-off value the document is saved in the database together with a (list of) subject classification(s). In the future this might be replaced with a more dynamic method that makes it easier to adopt to new termlists or changes in old ones. A modified version of this algorithm that takes into account word position in the text and proximity for boolean terms is now being tested.

The algorithm produces a list of suggested classifications and corresponding relevance scores using the algorithm:

$$\text{Relevance\_score} =$$

$$\sum_{\text{locs}} \left(\sum_{\text{terms}} (hits[\text{loc}_j][\text{term}_i] * weight[\text{term}_i] * weight[\text{loc}_j]) \right)$$

In order to take into account the hierarchical structure of the classification system experiments with score propagation are done, where scores are propagated to the most specific classifications suggested. This is achieved by assigning, for each leaf node in the list, the sum of it's score and all scores for all suggested classifications above in the classification tree, to that leaf node.

For those documents that are considered relevant with regard to the topic definition additional information is extracted to prepare them for further processing in the ALVIS system. The most important are character set normalization, language identification, and simple segmentation. Character set is determined by inspecting meta-tags bearing character set information (content-type). The document is also normalized in the sense that HTML tagging is cleaned up (using the Tidy library) and the character set is converted to UTF-8. Language is identified by a freely available language identification module.

Optionally topic specific RageRanks are calculated as a complement to the scores assigned by the algorithm above. These will be used for scheduling of URLs for crawling in order to obtain the most relevant pages as fast as possible.

The crawler stores all this information locally in a relational database that is also used for administration of the crawler. The crawler runs continuously in order to keep the topic-specific database as up to date as possible.

Presently we have test databases in five different topical areas: Carnivorous plants, Engineering, Materials Science, Malaria and Search Engines.

The crawled documents are exported to the ALVIS document processing pipeline. During export documents are converted into the ALVIS canonical XML format. In preparation for linguistic processing simple text segmentation is also done at this stage.

3.2 Acknowledgments

ALVIS is an EU Sixth Framework Programme (FP6), Information Society Technologies funded project (IST-1-002068-STP), started 2004-01-01 and will last for 3 years. Partners come from Finland, France, Switzerland, Sweden, Denmark, Spain, Slovenia, and China. Part of this material comes from internal reports.

4. REFERENCES

[1] ALVIS -Superpeer Semantic Search Engine. http://www.alvis.info/.

[2] Knowledge discovery and digital library research group. http://www.it.lth.se/knowlib/.

[3] K. Aberer, F. Klemm, M. Rajman, and J. Wu. An Architecture for P2P Information Retrieval. In *27th Annual International ACM SIGIR Conference (SIGIR 2004), Workshop on Peer-to-Peer Information Retrieval, Sheffield, UK*, July 2004.

[4] A. Ardö. Combine Web crawler, 2005. Software package for general and focused Web-crawling. http://combine.it.lth.se/.

[5] A. Ardö and T. Koch. Automatic classification applied to the full-text internet documents in a robot-generated subject index. In *Online Information 99, Proceedings*, pages 239–246, Dec. 1999. http://www.it.lth.se/anders/online99/.

[6] W.Buntine. Open Source Search: A Data Mining Platform. In *SIGIR Forum*, 2005. To appear.